Enhancing Distribution System Resilience: A First-Order Meta-RL algorithm for Critical Load Restoration

Zain ul Abdeen[1], Xiangyu Zhang[2], Waris Gill[1], Ming Jin[1]

I. Theoretical proofs

A. preliminaries

Before presenting the proofs of our main results, in this section, we establish necessary definitions and notational conventions. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function, the sub-gradient of f at a point x is denoted by $\partial f(x)$. We say that f is μ-strongly convex over a convex set $V \subseteq \text{int dom}(f)$ with respect to a norm $\|\cdot\|$ if, for any $x, y \in V$ and $g \in \partial f(x)$, it holds that $f(y) \geq f(x) + \langle g, y-x\rangle + \frac{\mu}{2}\|x-y\|^2$. Moreover, define $\psi : \mathcal{X} \to \mathbb{R}$ as a strictly convex and continuously differentiable function on int$\mathcal{X}$. The Bregman Divergence associated with ψ is given by $B_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle$, assuming ψ is strongly convex with respect to the norm $\|\cdot\|$ on int$\mathcal{X}$.

B. Proof of Task-Average-Optimality-Gap for Meta-based RL Algorithm

This section provides a detailed proof of the Task-Average-Optimality-Gap for our proposed meta-based RL algorithm. We begin with an analysis of the ES-RL algorithm applied to a single task, initially establishing a regret bound.

1) Single Task Analysis: Consider a series of Markov Decision Processes, where RL tasks emerge sequentially, indexed by $m = 1, \ldots, M$. In each task m, the agent refines its policy parameter $\{\phi_{m,j}\}_{j=0}^T$ over T iterations using the ES-RL algorithm. We present the following theorem with convergence guarantees for the ES-RL algorithm:

Theorem 1.1 (Theorem 6; [1]): If the ES-RL policy updates for each task m perform $T = \frac{4(N+4)^2L^2R^2}{\epsilon^2}$ iterations with a learning rate $\alpha_m = \frac{R}{(N+4)(T+1)^{1/2}L}$, and if $\sigma \leq \frac{\epsilon}{2L\sqrt{N}}$, then the sub-optimality gap for each task m is bounded by:

$$\mathbb{E}\left[F_m\left(\hat{\phi}_{m,T}\right)\right] - F_m\left(\phi_m^*\right) \leq \frac{2(N+4)L\|\phi_m^* - \phi_{m,0}\|}{\sqrt{T}}, \tag{1}$$

where, ϕ_m^* represents the parameters of the optimal policy π_m^*, and R bounds $\|\phi_m^* - \phi_{m,0}\| \leq R$.

2) Extension to Multiple Tasks: Extending the single task analysis to multiple tasks within the meta-learning framework, the task average optimality gap across multiple tasks within the meta-learning framework is defined as:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\left[\mathbb{E}\left[F_m(\hat{\phi}_m)\right] - F_m(\phi_m^*)\right] \\ \leq \frac{2(N+4)L}{M\sqrt{T}}\sum_{m=1}^{M}\|\phi_m^* - \phi_{m,0}\|.\end{aligned} \tag{2}$$

The right side of inequality (2) shows that the task-averaged regret is upper bounded by terms based on parameter initialization $\phi_{m,0}$. As the meta-algorithm sequentially updates these initial parameters through online learning, it is expected to reduce the task average sub-optimality as more tasks are addressed. We can consider the right-hand side of (2) as an individual loss function (i.e., $l_m(\phi_{m,0}) := \|\phi_m^* - \phi_{m,0}\|$), allowing us to bound the dynamic regrets (i.e., TAOG), measured by a dynamic sequence of optimal policy parameters $\{\phi_m^*\}_{m=1}^M$, via static regret, which is measured against a fixed policy parameter ϕ.

3) Static Regret Analysis: In this section we provide the static regret bound, which are used to furnish the upper bound on TAOG of the proposed algorithm. The lemma below provides a bound on the static regret:

Lemma 1.1 ([2]): Assuming the domain of the loss function is a non-empty closed convex set and the Bregman divergence is γ-Lipschitz continuous with $D_b = \max_{a,b\in\text{Dom}(f)} B_\psi(a, b)$, let η_m be a non-increasing sequence. Employing implicit online mirror descent or Follow The Regularized Leader (FTRL) on a sequence of loss functions $\{l_m\}_{m=1}^M$ where $l_m(\phi_{m,0}) = \|\phi_m^* - \phi_{m,0}\|$, the static regret against a fixed comparator ϕ_0^* is bounded by:

$$\frac{1}{M}\sum_{m=1}^{M} l_m(\phi_{m,0}) - l_m(\phi_0^*) \leq \frac{D_b}{\eta_m M} + \frac{\sum_{m=1}^{M}\delta_m}{M}, \tag{3}$$

where $\delta_m = l_m(\phi_{m,0}) - l_m(\phi_{m+1,0}) - \frac{B_\psi(\phi_{m+1,0},\phi_{m,0})}{\eta_m}$.

Theorem 1.2 (Theorem 6.2; [2]): Under the assumptions of Lemma 1.1, and if η_m is a decreasing sequence, the average static regret is bounded by:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M} l_m(\phi_{m,0}) - l_m(\phi_0^*) \leq \frac{2}{M}\min\{\sqrt{\beta\sum_{m=1}^{M}\mathbb{E}_m\left[g_m^2\right]}\}, \\ (l_1(\phi_{1,0}) - l_M(\phi_{M+1,0}) + V_M)\},\end{aligned} \tag{4}$$

where $V_M(f) = \sum_{m=2}^{M}\max_{\phi_{m,0}\in\text{Dom}(f)}|l_m(\phi_{m,0}) - l_m(\phi_{m-1,0})|$ is the temporal variability of the loss function.

[1]The Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA. Emails: {zabdeen,waris, jinming}@vt.edu

[2]National Renewable Energy Laboratory, Golden, CO, USA. Email:{Xiangyu.Zhang,}@nrel.gov

Note that the regret bound analyzed above is defined with respect to the optimal initial policy parameters ϕ_0^* in hindsight, not the final learned policy parameters.

C. Proof of Main Result

In Meta-RL, the extent to which TAOG improves is influenced by the similarity among the sequential MDP tasks [?]. For any fixed initial policies parameters $\{\phi\}$, the task similarity can be measured by $D^{*2} = \min_{\phi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \frac{1}{M} \sum_{m=1}^{M} \|\phi_m^* - \phi\|$. If the optimal policy parameter is not unique, we take the worst case for D^*, i.e., a set of policies for which D^{*2} is maximum. Building on the established foundations, we present a proof of the main theorem concerning the task average optimality gap:

Theorem 1.3 (Task Average Optimality Gap): Let $\{\phi_{m,0}\}_{m=0}^{M}$ be the initialization for each task determined by follow the average leader. For each task we train the policy for T steps with learning rate α and obtain $\left\{\hat{\phi}_{m,T}\right\}_{m=1}^{M}$. Let ϕ_m^* is the optimal Meta initialization for each task, then the task average optimality gap is bounded as

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[F_m(\hat{\phi}_{m,T})\right] - F_m(\phi_m^*) \leq \mathcal{O}\left(\frac{V_M + D^*}{\sqrt{T}M}\right). \quad (5)$$

Proof: Define

$$\bar{R} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[F_m(\hat{\phi}_m)\right] - F_m(\phi_m^*). \quad (6)$$

Given the bounds established in Theorem 1.1, TAOG is further bounded by:

$$\begin{aligned}
\bar{R} &\leq \sum_{m=1}^{M} \frac{2(N+4)L}{MT^{1/2}} (\|\phi_m^* - \phi_{m,0}\|) \\
&\overset{1}{=} \frac{2(N+4)L}{MT^{1/2}} \sum_{m=1}^{M} (\|\phi_m^* - \phi_{m,0}\|) - (\|\phi_m^* - \phi_0\|) + (\|\phi_m^* - \phi_0\|) \\
&\overset{2}{=} \frac{2(N+4)L}{MT^{1/2}} \sum_{m=1}^{M} (l_m(\phi_{m,0}) - l_m(\phi_0^*)) \\
&+ \frac{2(N+4)L}{MT^{1/2}} \sum_{m=1}^{M} (\|\phi_m^* - \phi_0\|) \\
&\overset{3}{\leq} \frac{2(N+4)L}{MT^{1/2}} (l_1(\phi_{1,0}) - l_M(\phi_{M+1,0}) + V_M) \\
&+ \frac{2(N+4)L}{MT^{1/2}} \sum_{m=1}^{M} (\|\phi_m^* - \phi_0\|) \\
&\overset{4}{\leq} \frac{2(N+4)L}{MT^{1/2}} (3D^* + V_M).
\end{aligned} \quad (7)$$

In the above proof inequality 3 is directly follows from the results in (4), and last inequality 4 follows from the definition of task similarity index. This completes the proof, showing the bounded nature of TAOG under our meta-learning framework. ■

References

[1] Nesterov, Yurii, and Vladimir Spokoiny. "Random gradient-free minimization of convex functions." Foundations of Computational Mathematics 17.2 (2017): 527-566.

[2] Campolongo, Nicolo, and Francesco Orabona. "Temporal variability in implicit online learning." Advances in neural information processing systems 33 (2020): 12377-12387.